Study on developing CERs with insufficient data

In Korean R&D environment

Sung Jin Kang, Yong Bok Lee, Dong Kyu Kim

Department of Operations Research, Korea National Defense University

CONTENTS

Background and Necessity (1/2)

◆ Current state of the cost analysis in Korea

Generally use the commercial parametric cost estimating models to analyze the cost of weapon system acquisition

Cross-checking the results of build-up vs. parametric method

◆ Limitations

- **Commercial models are developed based on foreign R&D data**
  - **Cannot validate the cost estimation results**
- **Cost items are different Korea defense industry accounting system from Output of commercial model**
  - **Cannot cross-check the results of two methods**

Background and Necessity (2/2)

◆ Need for development Korean version parametric cost estimating model

- **Suitable for our defense industry environment**
- **Estimate appropriate budget in early phase of acquisition process**
- **Develop using R&D, Production, O&S data in Korea**
- **Serve the estimates according to Korean defense industry cost accounting system**

CER Development_Methodology (1/2)

◆ Parametric cost estimating method

estimate cost using validated cost estimating relationships (CERs)
between projects known programmatic, technical, cost data

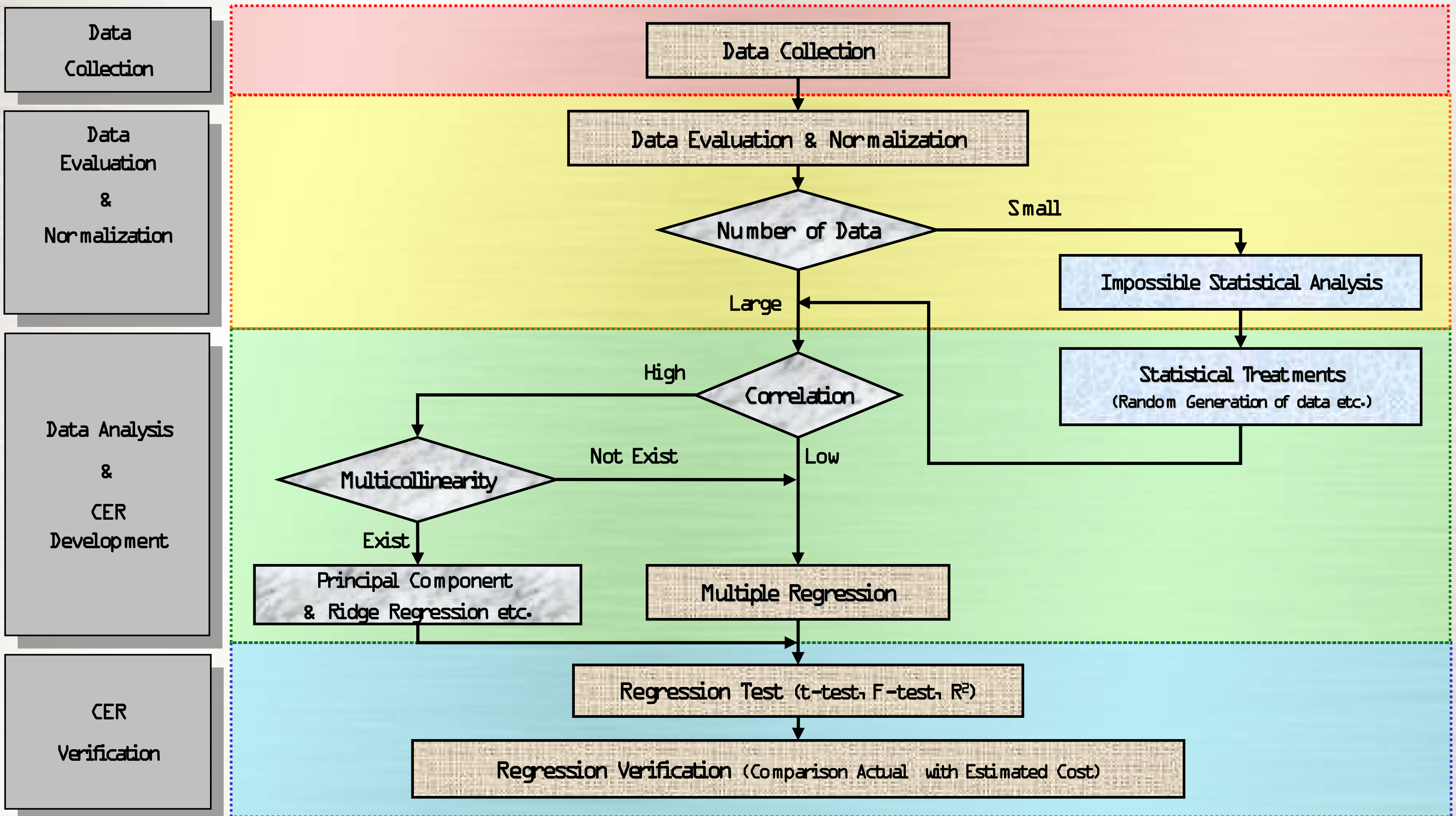
◆ Parametric Model is consist of many CERs

CERs : Mathematical expressions or formulas
that are used to estimate the cost as a function of
one or more independent variables

Establish the our own process of CER development

CER Development_Methodology (2/2)

◆ Procedure of CER development



CER Development_Data Collection

- Primary data Collection
  - 284 historical R&D data acquired in first level

    Collected from Agency for Defense Development

    Classify data into the 8 Categories according to Korea standard classification of weapon system

  □ first level : Communication and Control, Movement, Aircraft, fire, etc.

- Data Types
  - **Cost Data : 4 items (R&D Cost, Production Cost, Import Cost, etc.)**
  - **Specification Data : 17 items (Weight, Length, Velocity, etc.)**
  - **Project Data : 5 items (R&D duration, Quantity, Company, etc.)**

CER Development
_Data Evaluation & Normalization(1/4)

◆ Preliminary Evaluation

- **Evaluate the similarity and number of dada in same category**

  Analyze the data to develop CERs in second level

  - similarity is weak in 9 categories (Communication, Ammunition, etc.)

  - Insufficient number in 16 categories (Ship, Aircraft, etc.)

- **Determine the 2 categories are appropriate statistically**

  4 movement weapon systems

  - 2 tanks, 2 armed vehicles

  8 gun

  - 4 mortars, 4 howeitzers

CER Development _Data Evaluation & Normalization(2/4)

◆ Variable Selection

- **Dependent Variable : Total R&D Cost**
- **Independent Variables**

Select based on literature study, specialist interview

Movement (5 items)	Gun (7 items)
Combat Weight, No. of passenger, Engine power, Range, Max Velocity Dummy variable(Tank:1, Armed vehicle:0)	Max Range, Caliber, Weight, Length, Max rapidity, Continue rapidity, R&D duration

◆ Normalization

- **Cost Unit : 100M in 2009 constant, as Korea inflation index**
- **Sizing Units : Convert to metric system (meter scale ; km, kg...)**

CER Development _Data Evaluation & Normalization(3/4)

◆ Evaluation & normalization result

- **Movement (4 equipments, 6 independent variables)**

Type		Combat Weight (Ton)	Number of Passenger (Person)	Engine Power (hp)	Range (km)	Max Velocity (km/h)	Dummy Variable	R&D Cost (100M)
Tank	A	54.5	4	1200	400	60	1	511.08
	B	55	3	1500	450	70	1	2727.23
Armed Vehicle	C	13.2	12	280	480	74	0	169.33
	D	25	12	750	450	74	0	1054.55

- **Cannot apply multiple regression analysis**

4(no. of data point) - 6(no. of independent variables) < 2

- **Generate 80 random data with normal distribution considering correlation among independent variables to conduct multiple regression analysis**

CER Development _Data Evaluation & Normalization(4/4)

◆ Evaluation & normalization result

- **Gun (8 equipments, 7 independent variables)**

Type		Max Range (Km)	Caliber (mm)	Weight (kg)	Length (cm)	Max Rapidity (rounds/min)	Continue Rapidity (rounds/min)	R&D Cost (100M)
Mortar	C	3.59	60	18	99	30	20	18.203
	D	1.8	60	21	82	30	18	12.729
	E	6.473	81	41	155	30	11	35.255
	F	4.737	81	81	130	12	5	17.850
How eitzer	G	11.274	105	2,260	231	3	1	37.637
	H	14.7	105	2,650	392	5	2	27.069
	I	18	155	6,890	701	4	2	43.071
	J	18	155	25,000	912	4	1	74.074

- **Can conduct multiple regression analysis**

$$8(\text{no. of data point}) - 6(\text{no. of independent variables}) \geq 2$$

CER Development_Data Analysis(1/5)

◆ Correlation analysis about selected variables

Movement Equipment

	Combat Weight	No. of Passenger	Engine Power	Range	Max Velocity	Dummy Variable
Combat Weight	1					
No. of Passenger	-0.940	1				
Engine Power	0.956	-0.894	1			
Range	-0.505	0.382	-0.376	1		
Max Velocity	-0.513	0.455	-0.341	0.883	1	
Dummy Variable	0.933	-0.975	0.869	-0.309	-0.402	1

Gun

Be able to predict the existence of multicollinearity

	Max Range	Caliber	Weight	Length	Max Rapidity	Continue Rapidity
Max Range	1					
Caliber	0.958	1				
Weight	0.704	0.778	1			
Length	0.913	0.954	0.898	1		
Max Rapidity	-0.836	-0.803	-0.509	-0.690	1	
Continue Rapidity	-0.823	-0.799	-0.488	-0.653	0.945	1

CER Development_Data Analysis(2/5)

- VIF test to detect multicollinearity
  - **Multicollinearity ?**

    Independent variables are correlated among themselves

    The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the dependent variable and the independent variables.

✓ **The estimated standard deviations of regression coefficients become large**

□ **only imprecise information may be available about the individual true regression coefficients.**

✓ **Adding(deleting) a independent variable changes the regression coefficients**

□ **interpretation of the regression coefficients as measuring marginal effects is often unwarranted.**

CER Development_Data Analysis(3/5)

◆ VIF test to detect multicollinearity (Cont)

- **Variance Inflation Factor (VIF) ?**

  A highly useful diagnostic method to detect the present of multicollinearity

  VIF measures how much the variances of the estimated regression coefficients are inflated as compared to when the independent variables are not linearly related.

$$VIF_k = \frac{1}{(1 - R_k^2)}$$

where $k = 1, 2, \ldots, p-1$

R_k^2 : The coefficient of multiple determination when X_k is regressed on the *p-2* other X variable in the model

- **VIF_k is equal to 1 when $R_k^2 = 0$ and greater than 1 when $R_k^2 \neq 0$**
- **Maximum VIF in excess of 10 is frequently taken as an indication of the severe multicollinearity**

CER Development_Data Analysis(4/5)

◆ VIF test to detect multicollinearity (Cont)

- **4 severe multicollinearities exist in movement equipments**
- **All variables are correlated in gun equipments**

Movement

variable	VIF
intercept	0
Combat Weight	75.307
No. of Passenger	34.043
Engine Power	32.029
Range	7.726
Max Velocity	6.707
Dummy Variable	51.776

Gun

variable	VIF
Max Range	16.13
Caliber	53.632
Weight	15.152
Length	40
Max Rapidity	10.99
Continuous Rapidity	13.158

- **Cannot conduct the ordinary multiple regression**
- **We consider some remedial measure for serious multicollinearity that can be implemented with ordinary least squares.**

CER Development_Data Analysis(5/5)

◆ Methods for CER development in two cases

Movement R&D Cost

- **For the regression modeling, generate random data additionally**
- **To remedy multicollinearity, conduct the Principal Component Regression**

Gun R&D Cost

- **To get more accurate estimates and to overcome serious multicollinearity, conduct the Ridge Regression**

Variables are standardized, because the each scale of them is different

$$Z_i = \frac{(X_i - \mu_i)}{\sigma_i}, \ (mean = 0, \mathrm{var} = 1 \ of \ \ Z_i)$$

CER Development_Movement R&D Cost(1/5)

◆ Principal Component Regression(PCR) ?

- **A regression analysis to overcome problems with the highly correlated variables.**
- **In PCR instead of regression the independent variables on the dependent variable directly, the principal components of the independent variables are used.**
  - □ Principal component :
    composite index that are uncorrelated and capture much of the information contained in the independent variables.
- **Interpretation of PCR is available by re-expressing principal component into original variables.**

CER Development_Movement R&D Cost(2/5)

◆ Principal Component Regression(PCR) ? (Cont)

$$Y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \Lambda + \beta_p C_p + \varepsilon$$

$$C_k = a_{1k} X_1 + a_{2k} X_2 + \Lambda + a_{pk} X_p, k = 1, 2, \Lambda, p$$

$where: 1.\ Var(C_1) \geq Var(C_2) \geq \ldots \geq Var(C_p)$

$2.\ C_1 \sim C_p$ *are independent*

$3.\ \sum_{k=1}^{p} a_{1k} = 1$

- **Principal components with the highest variance are selected,**

  C_1 has the major variation within the all independent variables.

  C_2 has the second major information within the independent variables except C_1

  finally, C_p has the smallest information

- **Methods of principal component calculation**

  Use Variance–Covariance matrix (scales are same among variables)

  Use Correlation matrix (scales are not same among variables)

CER Development_Movement R&D Cost(3/5)

◆ Generate random data

- 80 random data are generated with normal distribution Considering correlation among independent variables

◆ Select the principal components using correlation matrix

- C_1, C_2 principal components are selected as proportion
- Dimension of data is reduced from 6 to 2

	Eigenvalue	Proportion	Cumulative
C_1	4.33838	0.7231	0.7231
C_2	1.34933	0.2249	0.9480
C_3	0.21312	0.0355	0.9835
C_4	0.06318	0.0105	0.9940
C_5	0.02941	0.0049	0.9989
C_6	0.00657	0.0011	1.0000

CER Development_Movement R&D Cost(4/5)

◆ Principal components calculation by eigenvector matrix

	Eigenvectors					
	C_1	C_2	C_3	C_4	C_5	C_6
Combat Weight	0.4700	0.1259	0.1984	0.1272	0.5274	-0.6555
No. of Passenger	-0.4554	-0.2123	0.3077	0.2838	0.6812	0.3290
Engine Power	0.4372	0.2516	0.5767	0.4223	-0.2624	0.4072
Range	-0.2959	0.6472	-0.4026	0.5668	-0.0016	-0.1010
Max Velocity	-0.3109	0.6207	0.4350	-0.5698	0.0478	-0.0441
Dummy Variable	0.4432	0.2676	-0.4267	-0.2811	0.4321	0.5331

$$C_1 = 0.4700Z_1 - 0.4554Z_2 + 0.4372Z_3 - 0.2959Z_4 - 0.3109Z_5 + 0.4432Z_6$$

$$C_2 = 0.1259Z_1 - 0.2123Z_2 + 0.2516Z_3 + 0.6472Z_4 + 0.6207Z_5 + 0.2676Z_6$$

- $Z_1 \sim Z_6$: standardized variables

◆ Regression equation with two principal components

$$R\ \&\ D\ \text{cost} = 1127.94 + 210.1C_1 + 567.15C_2$$

$$C_1 = 0.4700Z_1 - 0.4554Z_2 + 0.4372Z_3 - 0.2959Z_4 - 0.3109Z_5 + 0.4432Z_6$$

$(Z_1=0.95,\ Z_2=-0.86,\ Z_3=0.57,\ Z_4=-0.76,\ Z_5=-1.03,\ Z_6=0.99)$

Final Movement R&D CER

R&D Cost

= - 4,010.67 + 9.08(Combat Weight) – 50.80(No. of Passenger)

+ 0.50(Engine Power) + 5.25(Range) + 30.89(Max Velocity)

+ 486.60 (Dummy Variable)

CER Development_Gun R&D Cost(1/3)

- Ridge Regression(RR) ?
  - **A variant regression to remedy multicolliearity problems by modifying the method of least squares to allow biased estimators of the regression coefficients.**
  - **An estimator has only a small bias and is substantially more precise than an unbiased estimator.**

Use ridge estimator *b(k)*

$$b(k) = (X'X + kI)^{-1} X'Y, 0 < k < 1$$

$X'X$ the correlation matrix of the X variables

k : a biasing constant

$X'Y$ is the vector of coefficients of correlation between Y and each X

CER Development_Gun R&D Cost(2/3)

◆ Ridge Regression(RR) ?

Choice of Biasing Constant k

Ridge trace method

A simultaneous plot of $b_i(k)$ estimated ridge regression coefficients for different values of k where it is deemed $b_i(k)$ first become stable in the ridge trace, usually between 0 and 1

- But, need for many calculations and K is selected subjectively

Calculation method

$$k = \frac{p\hat{\sigma}^2}{\hat{b}'\hat{b}}$$

$\hat{b}$: a vector of least squares estimator

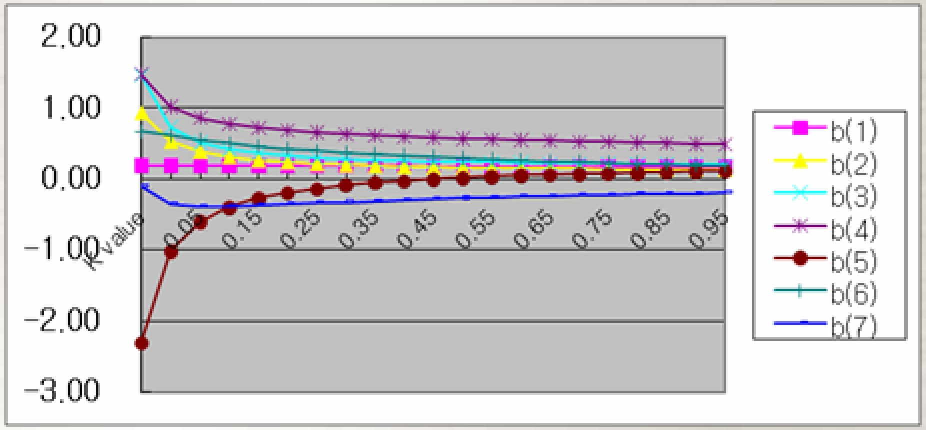
p : number of variables

$\hat{\sigma}^2$: residual mean square

- Easy calculation and k appear to have become reasonably stable

CER Development_Gun R&D Cost(3/3)

◆ Ridge trace and k calculation



$$k = \frac{p\hat{\sigma}^2}{\hat{b}'\hat{b}} = \frac{7 \times 2.67^2}{10.993} = 0.4129$$

Gun R&D CER

R&D Cost

= 0.615 + 0.562(Mat Range) + 0.131(Caliber)+0.001(Weight)

- 0.003(Length) + 0.527(Max Rapidity) – 0.758(Continuous Rapidity)

CER Verification

◆ Statistical Verification

(Movement R&D CER)

ANOVA

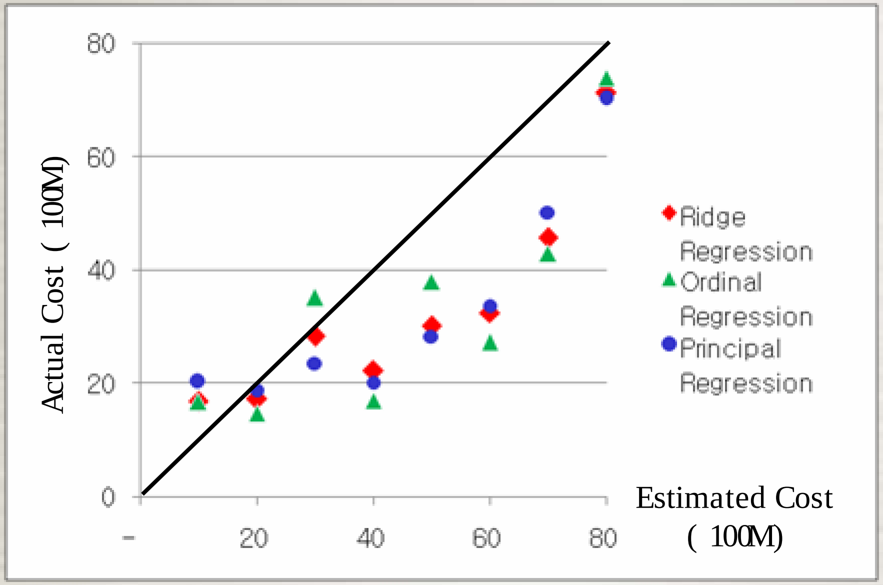
Source	DF	SS	MS	F	p
Model	2	49416851	24708425	61.13	0.0001
Error	77	31121572	404176		
Total	79	80538422			
R-Square=0.61, Adj. R-Square=0.60					

T-test

Variable	DF	Parameter	SE	T	p
Intercept	1	1127.94	71.08	15.87	<0.0001
C_1	1	210.10	34.34	6.12	<0.0001
C_2	1	567.15	61.58	9.21	<0.0001

◆ Graphical Verification

(Gun R&D CER)



- **Fitness of CERs is statistically satisfied**

  Predictive power(*Adj.* R^2*=0.6*) and use of Variables*(p-value<0.0001)* are appropriate

- **PCR and RR is more accurate than ordinal regression graphically**

Result and Future Study

- **Suggest a CERs development process considering the lack of data**
  - **Prevention against the loss of information through random data generating**
  - **PCR, RR approach to remedy multicollinearity in insufficient data environment**
  - **Verification of the CERs by statistical and graphical methods**
- **However, validation of CERs using same data family is impossible due to historical similar R&D case is not exist**

Continuous application of the appropriate regression methods considering data characteristic

Q & A

국방대학교

Thank you !